

Position Paper: OWL/RDF/LSID Utilization in NCI Cancer Research Infrastructure

Authors:

Frank Hartel
National Cancer Institute Center for Bioinformatics
hartelf@mail.nih.gov

Denise B. Warzel
National Cancer Institute Center for Bioinformatics
warzeld@mail.nih.gov

Peter Covitz
National Cancer Institute Center for Bioinformatics
covitzp@mail.nih.gov

Background

The National Cancer Institute Center for Bioinformatics, (NCICB) produces a public domain technology stack, caCORE^[i] <http://ncicb.nci.nih.gov/core>. NCI employs caCORE widely, building our numerous community portal sites and metadata and data repositories on it. Increasingly caCORE is being used to implement clinical and basic science resources at institutions external to NCI. caCORE is the infrastructure backbone supporting data management and application development at NCICB. caCORE has three primary components: (Enterprise Vocabulary Services (EVS), cancer Data Standards Repository (caDSR) metadata management, and cancer biomedical data "objects" (caBIO) implemented in an integrated software architecture. Effective with the caCORE 3.0 release, the object, metadata and ontology components of caCORE will share common base semantics.

Together the components of caCORE form the basis for interoperable analytic tools. NCICB provides APIs to facilitate access to these resources, documented in the caCORE Technical Guide available from the NCICB web site (<http://ncicb.nci.nih.gov/core>). All NCICB-developed caCORE components are distributed under open-source licenses that support unrestricted usage by both non-profit and commercial entities. caCORE is a heavyweight technology stack which is intended to be scalable and extensible, and suitable for use across the enterprise.

The NCICB looks to facilitate the identification and adoption of information technology and web standards to make these resources available to broad spectrum of cancer researchers and resources wherever they may reside. The NCICB recognizes the Semantic Web as a potent technology for meeting the needs of biomedical science and clinical care. Experience to date indicates that, within an enterprise, caCORE offers attractive capabilities although, like most heavy weight technologies, it requires commitment on the part of the adopting enterprise. However it does not currently adequately address the requirement for inter-enterprise integration in a way that is adequate to meet the needs of major collaborations such as caBIG <http://cabig.nci.nih.gov/>.

The cancer biomedical informatics grid, caBIG, is a partnership among the NCI-designated Cancer Centers and NCI to create a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools by leveraging caCORE. This platform will allow research groups to tap into the rich collection of emerging cancer research data and tools created to support cancer research while supporting their individual investigations. The caBIG architecture must support sharing across the Cancer Centers and NCI. Enterprise-wide sharing such as caCORE currently provides, is not enough. caBIG requires trans-enterprise sharing such as the Semantic Web or grid technology can provide.

Toward Integrated Semantics in Cancer Research

One of the major products of the caCORE is the NCI Thesaurus, a controlled terminology which exhibits ontology-like properties in its construction and use. The Thesaurus provides much of the base semantics to the other components of the caCORE. NCI Thesaurus is published in several formats (<ftp://ftp1.nci.nih.gov/pub/cacore/EVS/>), one of which is OWL Lite^[ii]. The decision to publish Thesaurus in Web Ontology Language (OWL) format

was made to make the Thesaurus available as a resource for agents and other Semantic Web components.

The future direction at NCICB is for the metadata to be shared using standard exchange formats and we are evaluating how best to integrate Resource Description Framework (RDF) metadata, ontologies, OWL and Life Sciences Identifiers (LSID) and other existing standards to support this goal. The caDSR of caCORE is based on the ISO 11179 metamodel for data element metadata. The metadata describe data elements used in cancer research. Some of the data elements are created and maintained manually using caDSR web based tools; some is derived from Object Management Group's (OMG) Universal Modeling Language (UML) class diagrams. UML class diagrams generated in caBIO provide a means for modeling various information including cancer biomedical entities, cancer data acquisition and analytic applications, and cancer data repositories and warehouses.

Briefly, all objects and attributes in the object space carry concept names and concept identifiers obtained from the EVS. Similarly all the administered components from which metadata entities are constricted in the caDSR carry concept names and concept identifiers obtained from the EVS. The EVS, caDSR and caBIO are described as UML models in the object space using UML and transformed into metadata in caDSR. All components are accessible through APIs. The UML models provide human readable descriptive representations of the data which also can be translated into metadata. By obtaining the words (lexical representation), order of the words (syntactic representation) and meaning (semantic representation) for describing the objects and metadata entries in caDSR using controlled terminology we have the ability to provide an immutable semantics that can be compared, interpreted and processed by computers.

UML class diagrams and controlled vocabularies form a powerful basis by which we can represent and transform information models into detailed semantically sound metadata. One missing piece of this infrastructure is the representation of object relationships in caDSR metadata, thus far documented in the UML and transformed only as reference information in caDSR, strings of text. Our next phase will include investigation of whether or not RDF and OWL can provide a standard means for transforming object relationships into caDSR metadata.

At the current time, the Semantic Web framework appears deficient with respect to tools that would enable us to access and transport the metadata structures of the caCORE. We have found that the semantics provided by terminology structures are not adequate, in and of themselves, to express the range of information that needs to be provided when sharing scientific and clinical data. Our solution is to build metadata based on ISO 11179, which is constructed from controlled terminology via the EVS, to express the range of information that must be provided if software systems are to be able to correctly find, interpret and use shared data. Hence the apparent inability of the Semantic Web as it exists today to surface metadata presented a road block.

NCICB is evaluating the adequacy of grid technology to meet the need to share data and applications and other resources among disparate enterprises. Our implementation, called caGRID, uses the GLOBUS Tool Kit and the Open Grid Services Architecture - Data Access Integration (OGSA-DAI) to enable interoperation, including sharing of data, across enterprises. Experience to date with caGRID suggests that this technology, along with our caCORE infrastructure, may be adequate to provide trans-enterprise interoperation while enabling components of the grid to effectively access metadata and terminology/ontology services needed to effectively share scientific data.

NCICB intends to test use of Semantic Web technology with caGRID. Like the caCORE stack, data grids are heavy; implementation and maintenance requirements will limit adoption. It may be possible to leverage Semantic Web technology to make access to, and use of, grid resources such as caGRID both easier and cheaper than they would otherwise be.

Currently access to the metadata is via caCORE APIs and caDSR tools which provide a means by which user can select all or a subset of data elements by downloading them in either of two predefined formats. There are two XML Schemas used for exporting caDSR metadata to end-users. These would potentially need to be modified to include RDF and OWL metadata. We believe that there are different purposes for "semantic" metadata along the lines of what RDF(S) is supposed to represent versus "system" (for lack of a better description) metadata that XML schema represents. Both were deemed to be necessary for the caBIG grid architecture.

One possible approach that we will probably test in the near term, is to expand this architecture to include RDF metadata and OWL metadata. It would be desirable for the W3C to specify a standard means for dealing with Registration Authority Identifiers (RAI) comprised of the Life Sciences Identifiers (LSID) which in turn would incorporate Object Identifiers (OID). There is a persistent need in bioscience to determine the reliability or quality of data. This need may be met to some degree by using LSID/OID as metadata to denote the owner or source of data.

Our next phase will include evaluating the addition of specialized metadata classes to our current caDSR infrastructure by comparing the metadata required for RDF and OWL to that needed to describe class behavior in UML as well as that needed to support the NCI Thesaurus. We anticipate that we will need to support XML Schema for system metadata and some kind of standard format for the semantic metadata, possibly RDF. There is obviously some overlap, so the appropriate separation and mappings of these two representations to each other will need to be sorted out. NCICB would welcome the opportunity to collaborate with W3C to define changes to the RDF/OWL/LSID specifications, if any, that our work may

indicate are desirable.

[i] Covitz, P.A., Hartel, F., Schaefer C., De Coronado, S., Fragoso, G., Sahni, H., Gustafson, S., and Buetow, K.H. "caCORE: A Common Infrastructure for Cancer Informatics", Bioinformatics, vol. 19 # 18, 2404-2412, 2003.

[ii] Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J. and Parsia, B. "The National Cancer Institute's Thesaurus and Ontology", Journal of Web Semantics, vol. 1, # 1, 75-80, 2003.